Nowcasting Inflation Using High Frequency Data

First version: June 2010
This version: January 2011

Michele Modugno[1]

Abstract

This paper proposes a methodology to nowcast and forecast inflation using data with sampling frequency higher than monthly. The nowcasting literature has been focused on GDP, typically using monthly indicators in order to produce an accurate estimate for the current and next quarter. This paper exploits data with weekly and daily frequency in order to produce more accurate estimates of inflation for the current and followings months. In particular, this paper uses the Weekly Oil Bulletin Price Statistics for the euro area, the Weekly Retail Gasoline and Diesel Prices for the US and daily World Market Prices of Raw Materials. The data are modeled as a trading day frequency factor model with missing observations in a state space representation. For the estimation we adopt the methodology exposed in Banbura and Modugno (2010). In contrast to other existing approaches, the methodology used in this paper has the advantage of modeling all data within a unified single framework that, nevertheless, allows one to produce forecasts of all variables involved. This offers the advantage of disentangling a model-based measure of "news" from each data release and subsequently to assess its impact on the forecast revision. The paper provides an illustrative example of this procedure. Overall, the results show that these data improve forecast accuracy over models that exploit data available only at monthly frequency for both countries.

Keywords: Factor Models, Forecasting, Inflation, Mixed Frequencies.

JEL classification: C53, E31, E37.

[1]European Central Bank and Universite Libre de Bruxelles, email mmodugno@ulb.ac.be

1 Introduction

Forecasting inflation is a concern for both market practitioners and central banks. Market practitioners tend to continuously update their expectations as new information is released, and to exploit this information in order to modify their investment strategies. Inflation is one of the variables that they continuously monitor. Central banks are charged with guaranteeing price stability, and therefore routinely monitor inflation expectations and forecasts. Namely, nowcasting inflation can help pinpoint the current inflation developments, understand the underlying forces that can jeopardise price stability, and thereby helping policy makers to recognise the need to take decisions that can offset these forces in a more timely manner.

HICP data in the euro area and CPI data in the US are usually released at the middle of the following month with respect to the reference month. A flash estimate for the euro area is available at the end of the reference month, but it only includes information on the total index. Nevertheless, during the month numerous data which carry valuable information on consumer prices, are released at weekly or daily sampling frequencies. This information, and the early signals that it contains, can be useful to improve the accuracy of the estimated actual inflation and its future developments.

This paper provides an econometric framework that allows interested parties to continuously update their inflation forecasts based on the growing amounts of information provided by all relevant available data. In order to achieve this, the paper uses two groups of data: First, the Weekly Commission Oil Bulletin Price Statistics (WOB) for the euro area and the Weekly Retail Gasoline and Diesel Prices (WRGDP) for the US. These sources contain surveys on the price at the pump of the fuels collected weekly, typically on Monday. In the euro area they are released in the following two or three days, while for the US they are released on the same day. Second, the World Market Prices of Raw Materials. They are published by the OCDE on the second or third day following the reference week, but the sampling frequency is daily.

These two sources contain information that can be very useful for forecasting inflation. Namely, the WOB and the WRGDP data are focused on consumer prices, which have, in comparison to raw oil price, the advantage that distribution and retail margins are fully accounted for. This is a desirable property in order to forecast consumer price inflation. Moreover, Raw Material Prices capture some of the global price dynamics as well as pricing information at the early part of the pricing chain. They can therefore provide an idea of fundamental price developments.

Although nowcasting inflation is a novel idea, there is a rather long literature focusing on nowcasting GDP. The use of higher frequency indicators in order to Nowcast/Forecast lower frequency indicators started with monthly data for GDP. GDP is a quarterly variable released with a substantial time delay (e.g. two months after the end of the reference quarter for the euro area GDP). In the meanwhile, several monthly indicators are released. Giannone, Reichlin, and Small (2008) for the US, Angelini, Bańbura, and Rünstler (2007) and Bańbura and Rünstler (2007) for the euro area, show that using monthly indicators is crucial in order to nowcast accurately GDP.

The methodology proposed in this paper fully exploits the co-movement of data with different frequencies. We model the data as a trading day frequency factor model with missing observations and we cast it in a state space representation.

The estimation adopts the methodology exposed in Banbura and Modugno (2010), which is in turn based on the methodology proposed by Watson and Engle (1983). Doz, Giannone and Reichlin (2006) show that the latter allows to estimate factor models by maximum likelihood

with large panels of data. Banbura and Modugno (2010) generalize their methodology in order to deal with panels with arbitrary patterns of missing data, e.g. when the dataset includes data sampled at different frequencies or with varying publication lags.

Two approaches have been proposed in order to use high frequency indicators for forecasting inflation. Lenza and Warmedinger (2010) average higher frequency (daily and weekly) data over a month and plug them in the dataset as monthly indicators for inflation. Alternatively, a new generation of models, the Mixed Data Sampling Regression Models (MIDAS), proposed originally by Ghysel, Santa-Clara, and Valkanov (2002) has been used by Monteforte and Moretti (2010) to forecast inflation in a two step approach. They extract principal components from a large sample of daily financial variables which are then used in the forecasting equation for the target variable. In order to prevent overparametrisation, the MIDAS approach assumes that the response to the high frequency explanatory variables follows a distributed lag polynomial.

In contrast to other procedures, e.g. MIDAS, the methodology proposed in Banbura and Modugno (2010) models data within a unified single framework that allows to forecast all the involved variables. This is achieved by using a factor model, which, by definition, does not suffer from overparametrisation. Moreover, the proposed methodology offers the possibility to disentangle model-based "news" from each release and to assess their impact on forecast revisions. Model-based "news" are defined as the difference between the released data and the respective model predictions. This is a desirable property, especially for central banks, because it gives the opportunity to identify the releases that modify the assessment on inflation, and to monitor whether releases that contain "news" can have potentially permanent effects on inflation. This information can be a significant input to monetary policy making. The paper provides an illustrative example of this procedure.

In order to asses the importance of using high frequency data for forecasting inflation this paper compares the forecast performance of the model using only monthly data, with the forecast performance of the model when also adding weekly and daily data. The provided empirical evidence shows that high frequency data are necessary in order to produce accurate inflation forecasts. Indeed the inclusion of this data not only improves forecasts at the shortest horizon, namely the nowcast, but even up to one year ahead. Looking at the forecasting performance of the subcomponents of inflation we can observe that most of the improvement comes from the more accurate forecast of the energy components in both the euro area and the US.

The paper is organized as follows: Section 2 describes the data, Section 3 introduces the model and the estimation methodology, Section 4 presents the results from the forecast exercise, Section 5 explains the concept of "news" and presents an illustrative example. Section 6 concludes.

2 Data

In order to produce an accurate nowcast of inflation we need to include in our analysis data that have two characteristics: high correlation with inflation and earlier availability with respect to the time release of inflation. Data with these two characteristics will help us to accurately track inflation in the current month. We focus on two groups of data:

The first group, contains energy prices for the euro area and the US from April 1996 to December 2009. As regards the euro area, data include the Weekly Commission Oil Bulletin Price Statistics (WOB). The Market Observatory for Energy presents consumer prices and net prices (excluding duties and taxes) of petroleum products in the euro area member states each week. These data are surveys of the fuel price at the pump, and they are released two or three days after the reference week by Eurostat. Our dataset includes the net prices, i.e. without duties and taxes, because of their longer availability. Similarly, as regards the US, the Energy Information Administration collects and publishes every Monday the Weekly Retail Gasoline and Diesel Prices (WRGDP), a survey of the pump prices for gasoline and diesel. These data include taxes and are the prices paid by the consumers. In comparison to raw oil price they both have the advantage that distribution and retail margins are fully accounted for.

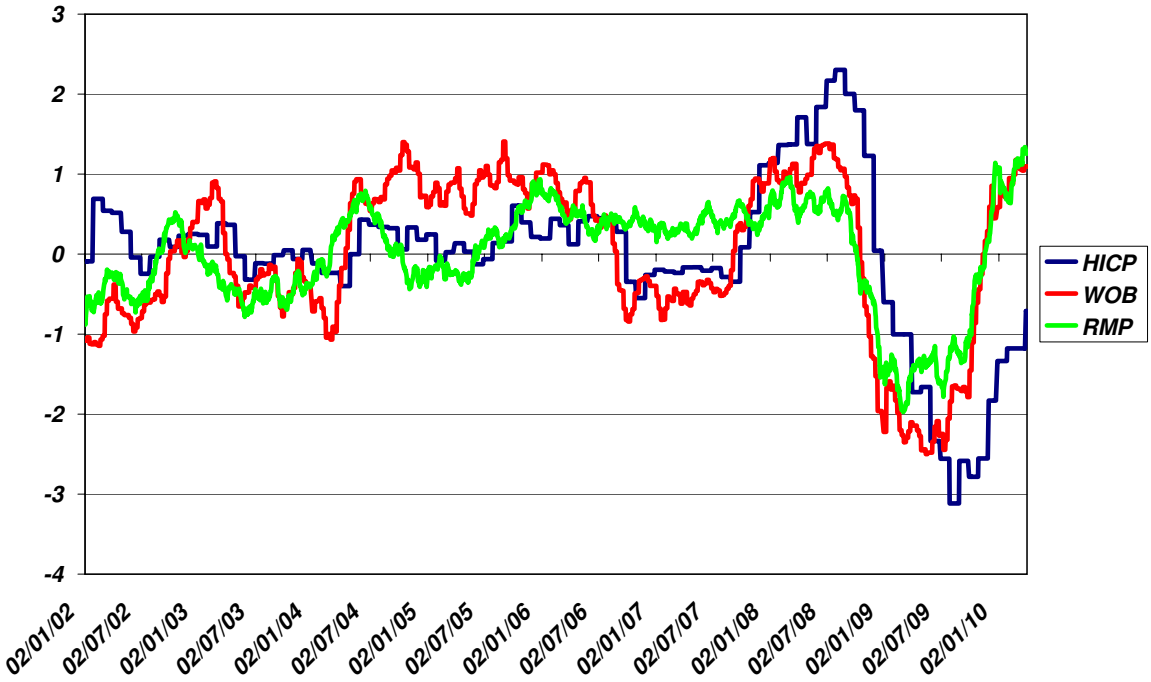
The second group of variables in our dataset includes the World Market Price of Raw Materials from April 1996 to December 2009. These data are sampled at daily frequency and are published weekly on the second or the third day of the following week compared to the reference week. They are produced by the Organisation for Economic Cooperation and Development (OECD) as weighted averages of the commodity imports of OECD countries. For the US we use these data expressed in dollars, for the euro area in euros. We include these series because they capture some of the global price dynamics as well as prices at the early part of the pricing chain.

Apart from the higher frequency data described above, the full dataset also comprises monthly series. We add to our euro area dataset 6 HICP series and to our US dataset 6 CPI series, i.e. respectively overall HICP and total CPI, our target variables, plus their components. They span the same period, i.e. from April 1996 to December 2009. We include those components in order to understand the effect of WOB/WRGDP and RMP data on the accuracy of our forecasts of the components. Thereby we also shed light on the underlying forces driving an improvement in the forecast accuracy of the target variables. The list of all the data is reported in Table 3 in Appendix 1.

The data employed co-move to a large degree with inflation data, as shown in Figures 3.1 and 3.2. Figure 1 shows the annual growth rates of the average of WOB data and raw material prices (RMP) compered to the annual growth rate of the overall HICP. As we can see these series can potentially be very informative in order to produce accurate nowcast of overall HICP inflation. Indeed these series are very much correlated with overall HICP inflation. Similarly, Figure 2 refers to the US and presents the annual growth rate of the average of WRGDP data and raw material prices (RMP) compared to the CPI annual inflation. As well as in the euro area, WRGDP and RMP series can be very informative for nowcasting and forecasting CPI inflation given their high correlation with total CPI.

Inflation data are considered to be timely, given that the first release in both the euro area and the US is published about 15 days after the reference month. Moreover, a flash estimate for the overall euro area HICP inflation is released the last day of the month. Until the last day, however, several data sampled at higher frequencies than monthly are already available. Therefore they contain more timely information about the current month. Among those we chose two groups of data that show high correlation with inflation.

Figure 1: Euro area data

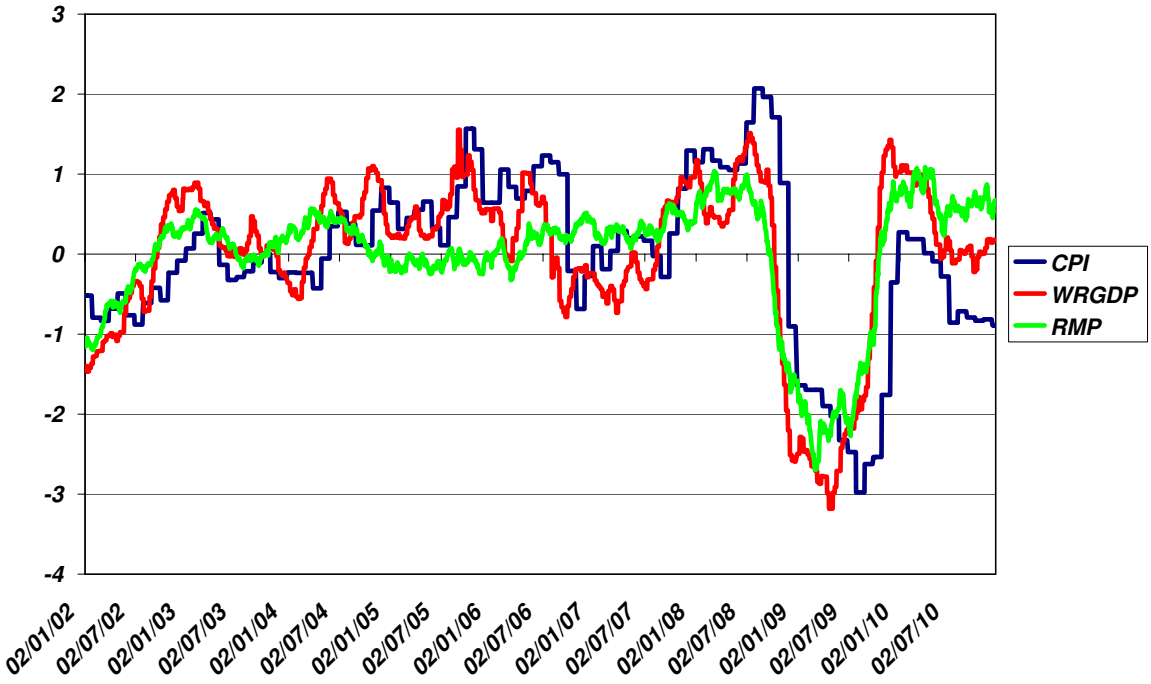


Notes: This Figure shows the evolution of the year on year growth rates of overall HICP, the average of the three series reported in the Weekly Commission Oil Bulletin (WOB) and the average of the raw material prices (RMP). All the series are standardized and centered to zero.

The timeliness of the higher frequency data employed can be further illustrated in Figure 3. It gives an example of the timing of the releases in March 2010 in the euro area (similar reasoning applies to the US). We begin by presenting the timing for the release of HICP data (upper part of Figure 3). The first information from HICP data for March 2010, i.e. the flash estimate for overall March HICP, was released only at the end of the month, i.e. 31 March. Before this day we do not have any information about March 2010 from HICP data. Earlier, on 16 March Eurostat published the first release of overall HICP and its component for the previous month, i.e. February 2010.

Now let us consider the lower part of Figure 3. During the current (reference) month, several WOB and RMP data are released. The first two releases, on 3 March for RMP and on 5 March for WOB, do not contain yet information about the current month, being RMP data relative to the trading days from the 22nd to the 26th of February, and WOB data relative to the week from the 23rd of February to the 1st of March. By the 12th of March data about the current month are released, i.e. on the 9th the RMP relative to the period from the 1st to the 5th of March, and on the 12th the WOB relative to the first week of March. Given the high correlation of WOB and RMP data with inflation, these data track inflation in March and

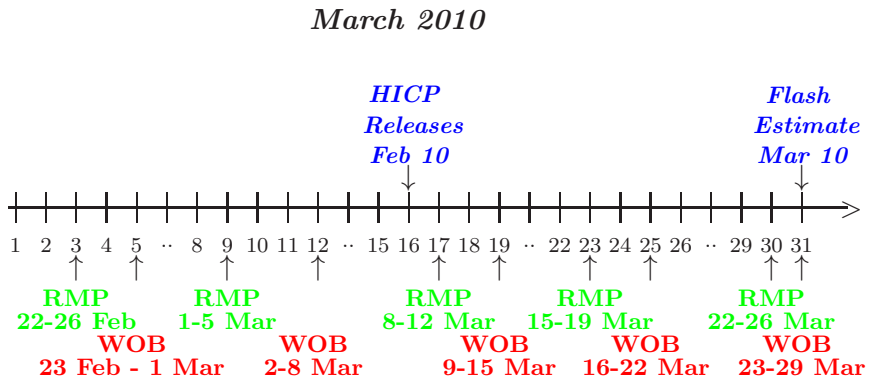
Figure 2: United States data



Notes: This Figure shows the evolution of the year on year growth rates of overall CPI, the average of the four series reported in the Weekly Retail Gasoline and Diesel Prices (WRGDP) and the average of the raw material prices (RMP). All the series are standardized and centered to zero.

provide information its future evolution, while still no information about the current month from HICP data is yet available. As time goes by more WOB and RMP data are released, and including them in our model will help to improve the accuracy of the nowcast. Moreover, as it is explained in Section 5, we can monitor which releases revise our nowcast and to which direction, giving the possibility to policymakers and market practitioners to understand and interpret the sources of revisions.

Figure 3: Timeliness



Notes: This Figure shows the flow of data released in a specific month, March 2010. The first HICP data on the current month (the flash estimate) is released on the last working day of the month. Before that, data from the Weekly Commission Oil Bulletin (WOB) and the World Market Price of Raw Materials (RMP) are released carrying information about the current month. Indeed on the 9th of March RMP data are released, covering the period from the 1st to the 5th of March, while on the 12th of March WOB data are released covering the period from the 2nd to the 8th of March, and so on.

3 Methodology

This paper adopts a dynamic factor model in order to produce forecasts of inflation. Factor models avoid overparametrisation summarising all data employed in few unobserved components which capture the correlation among the data. This allows us to exploit the information contained in many series contemporaneously.

3.1 Estimation

The general representation of a dynamic factor model is:

$$y_t = C f_t + \epsilon_t, , \qquad \epsilon_t \sim iid\,(0, \Sigma)\,, \tag{1}$$

where y_t is an $n \times 1$ vector of observations, C is an $n \times r$ matrix of loadings and f_t is an $r \times 1$ vector of unobserved components that display VAR dynamics:

$$f_t = A(L) f_{t-1} + u_t\,, \qquad u_t \sim iid\,(0, Q)\,, \tag{2}$$

while ϵ_t is an $n \times 1$ vector.

From now on t will refer to trading days. The general representation of a dynamic factor model can be estimated in several different ways. In this specific case we have to choose an estimation methodology that can deal with two complications: First, given that the vector y_t is composed by data observed at different frequencies, we observe missing data when modeling at the highest frequency available (in our case daily). Second, as we will show in subsection 3.2, we will need to impose restrictions on the parameters C, A, Q and $\Sigma = diag(\sigma_1, ..., \sigma_n)$.

In order to deal with these two issues we adopt the estimation methodology proposed in Banbura and Modugno (2010), that generalizes, for the case of missing data, the methodology proposed by Watson and Engle (1983). The latter methodology is based on the Expectation-Maximization (EM) algorithm under the assumption of an exact factor model, i.e. without serial and cross-correlation in the idiosyncratic components. Doz et all (2006) argue that these assumptions could be too restrictive. In particular, for the case of large cross-sections, they study the approximate factor model, allowing weak serial and cross-correlation in the idiosyncratic component. They show that as $n, T \to \infty$ the factors can be consistently estimated by quasi maximum likelihood, i.e. assuming that the model is a mis-specification of the exact factor model (see Doz, Giannone, and Reichlin, 2006, for the technical details). Consequently, the estimators are asymptotically valid also in the case of approximate factor models.

Banbura and Modugno (2010) also shows how to estimate the parameters in case of arbitrary pattern of missing data, e.g. when the dataset includes data sampled at different frequencies or with varying publication lags. It tackles this issue by deriving the parameters C, A, Q and Σ under the assumption that observations are defined as follows:

$$y_t = W_t y_t^{(1)} + (I_n - W_t) y_t^{(2)} \epsilon_t, \tag{3}$$

where W_t is a diagonal matrix of size n where the i^{th} diagonal element is equal to 0 if y_{it} is missing and equal to 1 otherwise, I_n is an identity matrix of dimension n and $y_t^{(1)}$ contains the non-missing observations at time t with 0 in place of the missing ones. This allows to impose a factor structure on the i^{th} variable only when y_{it} is available. Moreover, Banbura and Modugno (2010) show how to impose restrictions on the parameters, in order to impose a block structure to the factor model.

3.2 Econometric framework

As explained in Section 2, the n variables composing our dataset are divided in three groups characterized by different sampling frequencies: monthly (m), weekly (w) and daily (d). Let us define:

- $Y_t^{(m)}$ as the logarithm of the monthly series Y in month m and on day t. Between two consecutive releases there are k_m trading days, which can vary from 15 to 23, depending on the month m.
- $Y_t^{(w)}$ as the logarithm of the weekly series Y on week w and on day t. Between two consecutive releases there are k_w trading days, which are usually 5, if there are no bank holidays in week w.
- $Y_t^{(d)}$ as the logarithm of the daily series Y, observed on each day t[1].

Given these definitions for the log-levels, we derive:

- $y_t^{(m)} = (Y_t^{(m)} - Y_{t-k_m}^{(m)}) * 100$, i.e. the monthly growth rate for the series sampled at monthly frequency.

[1](d) in this case is used for illustrative purposes and in analogy with (m) and (w), but essentially refers to the same trading day t

- $y_t^{(w)} = (Y_t^{(w)} - Y_{t-k_w}^{(w)}) * 100$, i.e. the weekly growth rate for the series sampled at weekly frequency.
- $y_t^{(d)} = (Y_t^{(d)} - Y_{t-1}^{(d)}) * 100$, i.e. the daily growth rate for the series sampled at daily frequency.

Our vector of observable is then $y_t = [y_t^{(m)}, y_t^{(w)}, y_t^{(d)}]'$.

Given that the three groups of variables express three different measures, i.e. monthly, weekly and daily growth rates, we have to modify equations (1) and (2) in order to have a coherent model, where the unobserved components are expressed with the same measure of the specific series that they load. We chose the following representation:

$$\begin{bmatrix} y_t^{(m)} \\ y_t^{(w)} \\ y_t^{(d)} \end{bmatrix} = \begin{bmatrix} C_m & 0 & 0 \\ 0 & C_w & 0 \\ 0 & 0 & C_d \end{bmatrix} \begin{bmatrix} f_t^{(m)} \\ f_t^{(w)} \\ f_t^{(d)} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{(m)} \\ \varepsilon_t^{(w)} \\ \varepsilon_t^{(d)} \end{bmatrix}, \tag{4}$$

where the transition equation becomes:

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_t^{(m)} \\ f_t^{(w)} \\ f_t^{(d)} \end{bmatrix} = \begin{bmatrix} \Xi_t^{(m)} & 0 & 0 \\ 0 & \Xi_t^{(w)} & 0 \\ 0 & 0 & A \end{bmatrix} \begin{bmatrix} f_{(t-1)}^{(m)} \\ f_{(t-1)}^{(w)} \\ f_{(t-1)}^{(d)} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ u_t^{(d)} \end{bmatrix} \tag{5}$$

where C_m, C_w and C_d are the loadings for respectively monthly, weekly and daily variables. $f_t^{(m)}$,$f_t^{(w)}$ and $f_t^{(d)}$ are the monthly, weekly and daily factors. Ξ_t^w is a time varying coefficient equal to zero the day after each release of the weekly data and equal to one elsewhere. Ξ_t^m is equal to zero the day after each release of the monthly data and equal to one elsewhere. A is the matrix of the autoregressive coefficients for the daily factors (for an illustrative purpose we assume that $f_t^{(d)}$ is characterized by a VAR(1) dynamic). Once the model is written in this state space form, the methodology proposed in Banbura and Modugno (2010) can be applied in straightforward manner.

A more detailed view of the state space representation of the model can be informative. As described in Section 2, HICP and CPI indexes are collected around the 15th of each month. As such, they can be considered as a snapshot of prices around that day. In order to explain how this paper models snapshot variables observed at monthly frequency within a daily factor model, suppose that the monthly variables are sampled every day, and take their daily growth rate, $\hat{y}_t^{(m)} = (Y_t^{(m)} - Y_{t-1}^{(m)}) * 100$. The monthly growth rate can be derived the daily one by summing the daily growth rates from the first day after the previous month release to the day of the actual release:

$$\begin{aligned} y_t^{(m)} = \sum_{i=t-k_m+1}^{t} \hat{y}_i^{(m)} = (Y_t^{(m)} - Y_{t-1}^{(m)} + Y_{t-1}^{(m)} - Y_{t-2}^{(m)} + ... \\ ... + Y_{t-k_m+1}^{(m)} - Y_{t-k_m}^{(m)}) * 100 = (Y_t^{(m)} - Y_{t-k_m}^{(m)}) * 100 \end{aligned}$$

This implies that the monthly growth rate of a variable can be recovered summing its daily growth rates, if available. There are no daily growth rates for HICP/CPI data, but daily data that co-move with HICP/CPI can be used to extract a common factor at daily frequency, $f_t^{(d)}$. Given the availability of $f_t^{(d)}$, $y_t^{(m)}$ can be rewritten as:

$$y_t^{(m)} = \sum_{i=t-k_m+1}^{t} \hat{y}_i^{(m)} = C_m \sum_{i=t-k_m+1}^{t} f_i^{(d)} = C_m f_t^{(m)}$$

where $f_t^{(m)}$ is the sum of the daily factors ($f_t^{(d)}$) from day $t - k_m + 1$ to day t.

In order to aggregate the daily factors -and keeping in mind that Ξ_t^m is a time varying coefficient that is equal to zero the day after each release of the monthly data and equal to one elsewhere- we can define:

$$f_t^{(m)} = \Xi_t^m f_{t-1}^{(m)} + f_t^{(d)}. \tag{6}$$

This is the definition for $f_t^{(m)}$ expressed in equation (5). It means that on the first day after the previous month's release, and given that by definition $\Xi_{t-k_m}^m = 0$, equation (6) implies that $f_{t-k_m+1}^{(m)} = f_{t-k_m+1}^{(d)}$.

On the second day after the previous month's release, by definition $\Xi_{t-k_m+2}^m = 1$ and consequently $f_{t-k_m+2}^{(m)} = \Xi_{t-k_m+2}^m f_{t-k_m+1}^{(m)} + f_{t-k_m+2}^{(d)} = f_{t-k_m+1}^{(d)} + f_{t-k_m+2}^{(d)}$.

Iterating this sum for the remaining days between two consecutive releases of monthly data, on the day of the release for the current month (t), we have that $f_t^{(m)} = \Xi_t^m f_{t-1}^{(m)} + f_t^{(d)} = \sum_{i=t-k_m+1}^{t} f_i^{(d)}$.

The same explanation applies to the weekly data. Indeed both WOB and WRGDP are snapshot data.

4 Forecast exercise: Design and results

In this section we first explain the design of the forecast exercise. We present the competing models and describe the chosen process to evaluate forecast performances. We then present the results.

4.1 Forecast exercise design

In order to understand whether high frequency data can improve inflation forecast accuracy we compare the following models:

- a factor model that includes only monthly variables (Mon), estimated at monthly frequency.
- a factor model that includes all the variables (All), estimated at trading day frequency.
- a naïve random walk model (benchmark model).

All these models are estimated first with the euro area data and then with the US data. We adopt a recursive estimation scheme that, for the first evaluation, which covers the period from April 1996 to January 2001. For both countries the evaluation sample spans the period from January 2002 to December 2009. We evaluate the forecasts at, 0- (nowcast), 1-, 3-, 6- and 12-months ahead. The results are expressed as the ratio of the root mean squared forecast error (RMSFE) produced by the factor models to the RMSFE produced by the benchmark naïve model (random walk). We chose the random walk as benchmark because, according to the literature, e.g. Atkeson and Ohanian (2001), this is the best model for forecasting inflation, especially at the shortest horizons.

The data have been downloaded in February 2010. We conduct a pseudo real-time exercise meaning that, at each time we evaluate our models, we consider only the observations available at that time. This implies that we mimic the real-time availability of the data but we disregard their revisions.

The exercise can be better illustrated with the help of Figure 3 for the euro area. There we see that the information enters gradually in the estimation. The quality of the information content of each release affects the accuracy of the nowcast/forecast. Our model can be evaluated on any given day of the month. We show the results obtained the day after HICP and CPI data relative to the previous month are released. For example, for the euro area, in March 2010 (see Figure 3) we produce the forecasts on the 17th of March, the day after the HICP data of February are released. At that point in time we have already three releases of higher frequency data that contain information about March, two for RMP data and one for WOB data, available on 9, 17 March and on 12 March respectively. On the contrary, at that point in time, we do not have any HICP data relative to the current month, and the information available at monthly frequency is the HICP release about the previous month (February). Therefore, on the 17th, the only information available for the month that we nowcast (in the example March 2010) are WOB and RMP data.

4.2 Results

One of the most debated topics about factor models is how to choose the number of factors in equation (1) and the number of lags for the VAR in equation (2), especially for forecasting purposes. Several solutions have been proposed, but none of them reached a clear consensus. This is the reason why, in order to mimic a proper out-of-sample forecast exercise, we show a set of results obtained by averaging the performance of different parametrisations of the model, and for two different datasets. The different parametrisations involve RMSFEs averages produced by 24 different model specifications. These specifications include variation in the number of factors (1 or 2) and the number of lags (from 1 to 12). The first dataset uses all data available (All) and the second only monthly data (Mon), as described in subsection 3.2.

Table 1 shows the forecast performance of the euro area overall HICP inflation and its components, expressed in RMSFE ratios, as explained in section 3.4.1. If these ratios are equal to one, the models and the random walk have, in average, the same performance. A ratio below (above) unity suggests that the model under consideration outperforms (underperforms) the random walk.

As we can see from Table 1 WOB and RMP data improve the forecasting power of the model at any horizon, especially the nowcast, for total HICP. Indeed the RMSFE ratio for model *All* is 0.56 at horizon 0, i.e. the nowcast, while model *Mon* RMSFE ratio is 0.70. Most of the gain is due to the better forecasting performance of model *All* for the energy component and, in lesser extent, for the industrial goods component. For the other components, i.e.

Table 1: *Euro Area*

	Total		Energy		Proc. Food		Unproc. Food		Ind. Goods		Services	
	All	Mon	All	Mon	All	Mon	All	Mon	All	Mon	All	Mon
12	0.73	0.81	0.69	0.76	0.81	0.81	0.64	0.73	0.77	0.83	1.10	1.10
6	0.71	0.79	0.73	0.81	0.72	0.71	0.59	0.65	0.71	0.74	0.94	0.88
3	0.70	0.77	0.72	0.82	0.71	0.70	0.60	0.63	0.63	0.63	0.83	0.79
1	0.64	0.74	0.62	0.78	0.69	0.69	0.63	0.65	0.53	0.59	0.71	0.71
0	0.56	0.70	0.51	0.73	0.68	0.68	0.65	0.68	0.50	0.58	0.65	0.65

Notes: This Table shows the ratios of the average Root Mean Squared Forecast Errors (RMSFEs) produced by the factor model over the RMSFE produced by the Random Walk, for overall HICP inflation and its components. Those are averages of the RMSFE produced by factor models with 1 or 2 factors in equation (1) and from 1 to 12 lags in the equation (2). The two factor models differ in terms of datasets used. *MON* uses only the monthly data listed in Table 3. *ALL* uses all the available data.

services, processed and unprocessed food, these daily and weekly data do not improve the forecast accuracy, implying that they do not have any extra information content.

Table 2 shows the forecast performance for the US total CPI inflation at mid-month obtained with the two alternative datasets (All and Mon).

Table 2: *United States*

	Total		Energy		Food and Bev.		Housing		Goods and Serv.		Transport	
	All	Mon	All	Mon	All	Mon	All	Mon	All	Mon	All	Mon
12	0.69	0.78	0.70	0.74	0.81	0.88	0.78	0.88	1.55	1.33	0.70	0.77
6	0.71	0.86	0.69	0.80	0.73	0.75	0.77	0.82	1.28	1.16	0.73	0.83
3	0.64	0.82	0.63	0.80	0.71	0.70	0.76	0.81	1.18	1.11	0.65	0.84
1	0.49	0.75	0.44	0.73	0.71	0.71	0.71	0.76	1.06	1.05	0.48	0.78
0	0.38	0.67	0.35	0.63	0.73	0.70	0.67	0.67	0.85	0.79	0.35	0.68

Notes: This Table shows the ratios of the average Root Mean Squared Forecast Errors (RMSFEs) produced by the factor model over the RMSFE produced by the Random Walk, for overall CPI inflation and its components. Those are averages of the RMSFE produced by factor models with 1 or 2 factors in equation (1) and from 1 to 12 lags in the equation (2). The two factor models differ in terms of datasets used. *MON* uses only the monthly data listed in Table 3. *ALL* uses all the available data.

As we can see from Table 2 the inclusion of WRGDP and RMP data improves the forecast accuracy for total CPI for the US at all horizons, even more than for the euro area. Again the nowcast appears to be the most affected horizon, as the RMSFE ratio for model *All* is 0.38 at horizon 0, while model *Mon* RMSFE ratio is 0.67. Also for the US, most of the gain is due to the better forecasting performance of model *All* for the energy component and, to a lesser extent, for the transport component. For the other components, i.e. food and beverages, goods and services and housing, including in the dataset these daily and weekly data does not significantly improve the forecast accuracy.

5 News and forecast revisions

Section 4 shows the performances of our model assuming that forecasts are produced once per month, i.e. after the HICP/CPI releases. However, when forecasting in real-time, continues inflow of information occurs as new figures for various predictors are released non-synchronously and with different degrees of delay. Therefore, in such applications, we seldom perform a single prediction for the reference period but rather a sequence of forecasts, which are updated when new data arrive. Intuitively, only the news or the unexpected components from the data released should revise the forecast, hence, extracting the news and linking it to the resulting forecast revision is key for understanding and interpreting the latter. The model employed, in contrast to models that consider high frequency data as predetermined regressors (e.g. MIDAS), has the advantage of treating all the variables as endogenous. This implies that our unique framework produces forecasts for all the variables and therefore it gives us the possibility to extract, for each variable, a model based "unexpected" component once new figures are released. This section shows first how the "unexpected" content, i.e. the "news", in a data release is linked to the resulting forecast revision. It then describes how the inflow of new information affected the nowcast of overall HICP inflation in March 2009.

5.1 *News* and forecast revision

Let Ω_{v-1} and Ω_v be two consecutive vintages of data, consequently $\Omega_{v-1} \subset \Omega_v$.[2] Let I_v denote the *news* in Ω_v with respect to Ω_{v-1}. For example, let us assume that the difference between Ω_{v-1} and Ω_v is the release of RMP data for the period t_i. The *news* is $I_v = y_{t_i}^{RMP} - \mathbb{P}\left(y_{t_i}^{RMP}|\Omega_{v-1}\right)$, where $y_{t_i}^{RMP}$ is a vector containing the last figures released.

Assume that we are interested in how "news" revises overall HICP inflation forecast for the period t_j. As $I_v \perp \Omega_{v-1}$ we can write

$$\mathbb{P}\left(\cdot|\Omega_v\right) = \mathbb{P}\left(\cdot|\Omega_{v-1}\right) + \mathbb{P}\left(\cdot|I_v\right)$$

or

$$\underbrace{\mathbb{P}\left(y_{t_j}^{HICP}|\Omega_v\right)}_{\text{new forecast}} = \underbrace{\mathbb{P}\left(y_{t_j}^{HICP}|\Omega_{v-1}\right)}_{\text{old forecast}} + \mathbb{P}\Big(y_{t_j}^{HICP}|\underbrace{I_v}_{\text{news}}\Big)\,.$$

In other words, the updated forecast can be decomposed into the sum of the old forecast and of the contribution from the *news* in the latest release.

To compute the latter we use the fact that

$$\mathbb{P}\left(y_{t_j}^{HICP}|I_v\right) = \mathrm{E}\left(y_{t_j}^{HICP} I_v'\right)\mathrm{E}\left(I_v I_v'\right)^{-1} I_v\,.$$

Furthermore, given equation (1) we can write

$$\begin{aligned} y_{t_j}^{HICP} &= C_{HICP} f_{t_j} + \epsilon_{t_j}^{HICP}\,, \\ I_v &= y_{t_i}^{RMP} - y_{t_i|\Omega_{v-1}}^{RMP} = C_{RMP}\left(f_{t_i} - f_{t_i|\Omega_{v-1}}\right) + \epsilon_{t_i}^{RMP}\,, \end{aligned}$$

where C_{HICP} and C_{RMP} are the rows of C corresponding to HICP and RMP, respectively. It can be shown (see Banbura and Modugno (2010)) that:

$$\begin{aligned} \mathrm{E}\left(y_{t_j}^{HICP} I_v'\right) &= C_{HICP}\mathrm{E}(f_{t_j} - f_{t_j|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})' C_{RMP}' \quad \text{and} \\ \mathrm{E}\left(I_v I_v'\right) &= C_{RMP}\mathrm{E}(f_{t_i} - f_{t_i|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})' C_{RMP}' + \Sigma_{RMP}\,, \end{aligned}$$

[2] In what follows, we do not take into account data revisions and changes in the parameter estimates. The influence of those factors needs to be analyzed separately.

where Σ_{RMP} is a diagonal matrix with elements of Σ corresponding to the RMP data. The expectations $\mathrm{E}(f_{t_j} - f_{t_j|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'$ and $\mathrm{E}(f_{t_i} - f_{t_i|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'$ can be obtained from the Kalman filter.

Consequently, we can find a vector B such that the following holds:

$$\underbrace{y^{HICP}_{t_j|\Omega_v}}_{\text{new forecast}} = \underbrace{y^{HICP}_{t_j|\Omega_{v-1}}}_{\text{old forecast}} + B\Big(\underbrace{y^{RMP}_{t_i} - y^{RMP}_{t_i|\Omega_{v-1}}}_{\text{news}}\Big). \tag{7}$$

This enables us to trace the sources of forecast revisions. More precisely, in the case of a simultaneous releases of several (groups of) variables it is possible to decompose the resulting forecast revision into contributions from the "news" in individual (groups of) series.[3] In addition, we can produce statements like e.g. "after the release of Raw Material Prices, the forecast of HICP inflation went up because the indicators turned out to be (on average) higher than expected".[4]

5.2 Tracking forecast revisions: March 2009

Figure 4 shows how our tool can be used to track forecast revisions. It displays how the nowcast of the euro area HICP inflation evolved in March 2009 when new figures were released.

In this Figure we show all relevant data in the progress of updating of nowcast as new information is been incorporated. The overall HICP inflation data appears as a series daily observations with a clear break after 16 March (from 1 to 16 March the series depicts the flash estimate and the first release data are shown afterwards). This is effectively the nowcast produced by a random walk model. These variables are measured on the left-hand vertical axis.

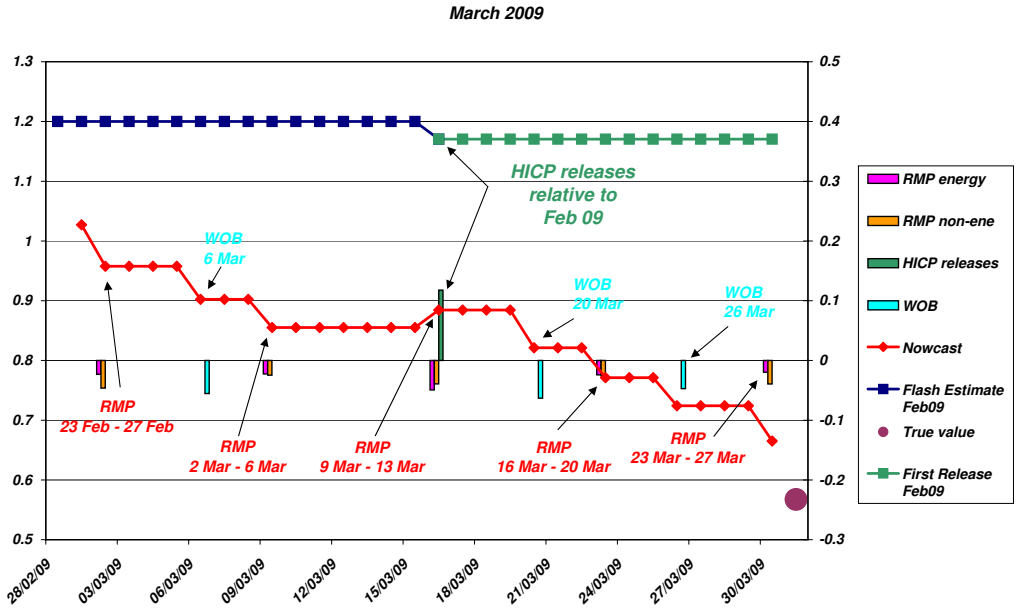
The nowcast from the model shows a more varied behaviour. Whenever new data are released, our nowcast is revised incorporating new information. The impact of the "news" on inflation (the coefficient B times the difference between the data and its expected value, as defined in equation (7)) for each group are measured on the vertical right-hand axes. We group RMP data in energy and non-energy (non-ene) and present the impact of these two groups as vertical bars. The energy group is composed by crude oil and coal, the non-energy one contains all the remaining series in Table (3).

As we can see, we can disentangle the effect of new releases on the nowcast revisions and indicate, in a model based framework, which given release produces a certain effect and quantify this effect. For example, on 2 March RMP data are released. Our nowcast for March 2009 is revised down, getting closer to the true value, and making our forecast more accurate. This happens because the data released for the two groups of variables, the energy and the non-energy, are lower than what the model was predicting. On 6 March the WOB data are released. Again the released data are lower than what the model was predicting, therefore the nowcast is revised downwards. The same happens on 9 March. On 16 March the nowcast is revised upwards, because of several new data releases. Namely the HICP first estimates of the different components relative to February 2009 and RMP data relative to the period from 9 to 13 March arrive. As it can be seen, both energy and non-energy raw material prices push the nowcast downwards, but not enough to compensate the effect of the HICP releases. These releases were

[3] If the release concerns only one group or one series, the contribution of its "news" is simply equal to the change in the forecast.

[4] This holds of course for the indicators with positive entries in B.

Figure 4: News



Notes: This Figure shows on the left-hand vertical axes the nowcast of overall HICP inflation for March 2009, and the nowcast produced by a Random Walk, i.e. the Flash Estimate until 16th when the First Estimate is released and becomes the new Random Walk nowcast. Every time new data are released, WOB data, RMP data or HICP first estimates of February 2009, our nowcast is revised. We distinguish between energy and non-energy (non-ene) components of RMP data, WOB data and HICP releases. The contributions of those releases to the nowcast revisions are measured on the right-hand vertical axes.

expected, from the model, to be much lower than their final realizations, bringing the nowcast far away from the true value. Clearly these data, given that they contain information about the previous month, are feeding the model with backward looking information that cannot help to improve the understanding of current inflation dynamics. As time proceeds, new RMP and WOB data are released, and the nowcast is again revised down getting closer and closer to the true value.

6 Summary

This paper proposes an econometric framework that exploits weekly and daily data in order to forecast the euro area overall HICP inflation and the US total CPI inflation.

The paper focuses on two groups of data with sampling frequency higher than monthly: the first is composed by the World Market Price of Raw Materials at daily frequency. The second includes weekly surveys of fuel prices at the pump, the Weekly Oil Bulletin Price Statistics for

the euro area and the Weekly Retail Gasoline and Diesel Prices for the US. We focus on these data because of two reasons: first they co-move with inflation in the respective areas (euro area and US), and second because they they become available in a more timely fashion compared to inflation data. To these two groups of data we add our target variables, i.e. the euro area overall HICP inflation and the US total CPI inflation, and their main subcomponents.

We model these data as a dynamic factor model estimated at daily frequency. The daily factors are aggregated in order to build weekly and monthly factors that explain, respectively, weekly and monthly data. This aggregation is obtained imposing time-varying coefficients in the state space representation of our model.

The estimation methodology is the one proposed in Banbura and Modugno (2010), that allows to estimate factor models on datasets with arbitrary patterns of missing data, e.g. with series sampled at different frequencies or with varying publication lags. Moreover, it allows us to introduce restrictions on the coefficients, such as the time-varying coefficients that we use in order to aggregate the daily factors. This framework offers original estimation advantages compared to the previous literature. It avoids the need of averaging high frequency data in order to obtain monthly frequency indicators, like in Lenza and Warmedinger (2010). This allows us to fully exploit the co-movement in our dataset, without losing any information. Moreover, in contrast to MIDAS models, like the one proposed in Monteforte and Moretti (2010), the proposed framework does not require to impose high frequency data as predetermined variables. It therefore allows to disentangle model-based "news" from each release and then to assess their impact on forecast revisions.

The results suggest that the chosen weekly and daily data are important to improve the forecast accuracy for both the euro area overall HICP and the US total CPI inflation, especially at the shortest horizon, i.e. the current month. The model performs considerably better compared to a model that uses only monthly information or compared to a random walk. This is especially due to the improved forecast accuracy of the energy components, which are the most volatile among the US CPI and euro area HICP components. Moreover, it is presented that illustrates the use of the model in identifying "news" effects (i.e. the revisions of the forecast of the target variable that arise from new data releases), further emphasizing the potential of this framework as an important tool for policy analysis.

References

Angelini, E., M. Bańbura, and G. Rünstler (2007): "Estimating and forecasting the euro area monthly national accounts from a dynamic factor model.," Discussion paper, European Central Bank.

Atkeson, A., and L. Ohanian (2001): "Are Phillips curves useful for forecasting inflation?," *Federal Reserve Bank of Minneapolis, Quarterly Review*, 25, 2–11.

Banbura, M., and M. Modugno (2010): "Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data," Working Paper Series 1189, European Central Bank.

Bańbura, M., and G. Rünstler (2007): "A look into the factor model black box. Publication lags and the role of hard and soft data in forecasting GDP.," Working Paper Series 751, European Central Bank.

Doz, C., D. Giannone, and L. Reichlin (2006): "A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models," Working Paper Series 674, European Central Bank.

Ghysel, E., P. Santa-Clara, and R. Valkanov (2002): "The MIDAS Touch: Mixed Data Sampling Regression Models," Working papers, UNC and UCLA.

Giannone, D., L. Reichlin, and D. Small (2008): "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, 55(4), 665–676.

Lenza, M., and T. Warmedinger (2010): "A Factor Model for Euro-area Short-term Inflation Analysis," mimeo ECB.

Monteforte, L., and G. Moretti (2010): "Real time forecasts of inflation: the role of financial variables," Working Paper Series 81, LUISS.

Table 3: Data

name	source
Euro area	
monthly	
HICP - Overall index	Eurostat
HICP - Processed food incl. alcohol and tobacco	Eurostat
HICP - Unprocessed food	Eurostat
HICP - Industrial goods excluding energy	Eurostat
HICP - Energy	Eurostat
HICP - Services	Eurostat
weekly	
Diesel	Eurostat
Euro Super 95	Eurostat
Gas Oil	Eurostat
United States	
monthly	
CPI - All items	Bureau of Labor Statistics
CPI - Energy	Bureau of Labor Statistics
CPI - Food and Beverages	Bureau of Labor Statistics
CPI - Housing	Bureau of Labor Statistics
CPI - Other goods and services	Bureau of Labor Statistics
CPI - Transportation	Bureau of Labor Statistics
weekly	
Diesel Sales Price	Energy Information Administration
Midgrade All Formulations Gas Price	Energy Information Administration
Premium All Formulations Gas Price	Energy Information Administration
Regular All Formulations Gas Price	Energy Information Administration
Euro area and United States	
daily	
Food and tropical beverages	Org. for Economic Cooperation and Development
Cereals	Org. for Economic Cooperation and Development
Oilseeds & oil	Org. for Economic Cooperation and Development
Beverages, sugar & tobacco	Org. for Economic Cooperation and Development
Industrial raw materials	Org. for Economic Cooperation and Development
Agricultural raw materials	Org. for Economic Cooperation and Development
Spinning material	Org. for Economic Cooperation and Development
Non-ferrous metals	Org. for Economic Cooperation and Development
Iron ore, scrap	Org. for Economic Cooperation and Development
Coal	Org. for Economic Cooperation and Development
Crude oil	Org. for Economic Cooperation and Development